DOCUMENT RESUME

ED 282 412 FL 016 703

AUTHOR de Jong, John H. A. L.
TITLE Listening, a Single Trait in First and Second Language Learning.
PUB DATE Jan 84
NOTE 16p.
PUB TYPE Reports - Research/Technical (143) -- Journal Articles (080)
JOURNAL CIT Toegepaste taalwetenschap in artikelen 20; n3 p66-79 Jan 1984

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Age Differences; Comparative Analysis; Correlation; Educational Background; *English (Second Language); *Language Processing; Language Skills; Language Tests; *Listening Comprehension; Native Speakers; *Second Language Learning; Statistical Analysis; Test Length; Test Reliability; *Test Validity

ABSTRACT

A study investigated the validity of an English listening skills test by comparing the results of native American and British English speakers with those of Dutch students of English as a second language. A hypothesis suggested that two-thirds of the items would test listening skills and the remaining third would test other knowledge. Test results were analyzed according to both classical test theory and the Rasch item response theory. The findings showed the test to be disappointing as a measure of listening comprehension skills, but did suggest that the language listening ability of first- and second-language learners can be measured along a single variable that can be distinguished from age and educational background. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED282412

LISTENING, A SINGLE TRAIT IN FIRST AND SECOND LANGUAGE LEARNING.

John H.A.L. de Jong
National Institute of Educational Measurement.
Cito, Arnhem.

REPRINT (pp. 66-79) FROM:

toegepaste taalwetenschap in artikelen 20

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

de Jong

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

jaargang 1984 **nummer 3**

VU boekhandel/uitgeverij
1007 MC AMSTERDAM - P.O. BOX 7161
THE NETHERLANDS - TEL. 020-548 26 54

FL016703

BEST COPY AVAILABLE

80.6256.4962

2

LISTENING, A SINGLE TRAIT IN FIRST AND SECOND LANGUAGE LEARNING.

John H.A.L. de Jong
National Institute of Educational Measurement.
Cito, Arnhem.

Introduction

In applied linguistics the pendulum regularly swings from theories based on a clear distinction between first and second language learning to theories stressing the similarities in both processes of language acquisition. The contrastive analysis hypothesis (Lado, 1957) relates learner difficulty to differences between target and native language. Language transfer in this theory is a dominant force in foreign language learning. In parallel with Chomsky's (1959, 1968) rejection of the structuralistic and behaviourist approach to language acquisition, the contrastive analysis hypothesis in its strong form proved to be untenable and theories on language learning have focused on understanding the principles of first language acquisition and their applicability to foreign language learning. The development in Corder's publications reflects this shift in attention (Corder, 1981). Krashen formulated a theory on the Monitor Model and language acquisition (Krashen, 1981, 1982; Burt e.a., 1982) in which the natural order hypothesis is clearly related to Chomsky's concept of an innate universal grammar (Chomsky, 1981). But Krashen's attempt to build a theoretical framework from a number of widely accepted ideas on second language acquisition lacks sufficient foundation (McLaughlin , 1978; Gregg, 1984; Corder, 1984) and language transfer is receiving renewed interest from applied linguists (Kellerman, 1983; Schachter, 1983). The swinging of the pendulum, however, causes the hands of the clock to move on. Gass (1984) postulates that language universals serve as an overall guiding principle in second language acquisition, interacting with the systems in the native and in the target language, thus combining both principles.

In language testing there is a controversy between advocates of discrete point testing and integrative testing. In the early years of testing the stress put on the necessity to break down language competence into different skills and even constituent components of these skills reflected the structuralist approach to language learning. In the "Post Modern Phase" (Spolsky, 1984) more attention has been given to the testing of language use, in 'authentic' situations, testing communicative competence (Carrol, 1980; Canale and Swain, 1980; Canale, 1983; Morrow, 1978), and at the same time holistic (Conlan, 1983) and impressionistic (Aghbar, 1983) scoring have come back into favour.

If first and second language acquisition are related, then the difficulties in processing samples of a language should be

similar for foreign language learners and for native speakers. This study is an attempt to prove that a single trait underlies the performance of native and non-native speakers on a listening comprehension test. The power of a test to measure differences in ability amongst individuals or groups of individuals depends on the homogeneity and validity of the set of items contained in the test and on the level of difficulty of the items in relation to the ability of the persons to be measured. The hypothesis to be tested then is: a set of items that discriminates amongst non-native speakers with respect to their level of ability in performing a particular foreign language task will discriminate on the same trait amongst native speakers of that language provided that the test is not too easy for them. De Jong (1983) demonstrated a procedure for making a best selection of items by means of a series of Rasch analyses. From a listening comprehension test badly fitting items with low discrimination indexes were deleted in each subsequent analysis. It was concluded that a selection of two thirds of the items in the test constituted a valid measure for listening comprehension of English as a foreign language. The remaining part was thought to test a different ability, possibly general intelligence or knowledge of the world. If the hypothesis is not rejected the same selection of items will discriminate consistently between native speakers differing in age and/or educational background and consequently differing in command of their mother tongue. A test composed of the rejected items, however, will reveal a different relation between the native speakers concerned, as this test taps a different trait.

Method

The test used in this study was constructed as a pilot test of listening comprehension of English as a foreign language at the Dutch National Institute for Educational Measurement (Cito, Arnhem) in a research project designed to develop new methods of testing listening comprehension (De Jong and Van den Nieuwenhof, 1982; De Jong, 1984). The test uses life recordings taken from British and American radio programmes cut into samples of about 20 seconds each. Testees listen to the tape once and have to respond to a multiple choice question with two options printed in a test booklet within the 10 second pause in between samples provided on the tape. Two item formats were used: true-false items (: was the statement in the test booklet in accordance with what was said on the tape?); and modified cloze items: words to be deleted from the text were chosen for their semantic relevance in the context. In each sample one word - or group of words - was cut out from the tape and replaced by an electronical sound. Testees had to decide which of the two options presented in their test booklet could be used to restore the text. The test in this study contained three types of language use: a discussion, a telephone

conversation and a regular news programme. Total test length was 59 items. The trait to be measured with the test was defined as: "The ability to understand the foreign language at the level of native speakers of comparable age and educational background" (De Jong, 1983). The test was administered to three groups of subjects:

1 A group of 30 native speakers of English, about 17 years old, taking A-levels in the British school in the Netherlands.
2 A group of 44 native speakers of English, from the American school in the Hague, 15 to 16 years old, two years from graduating at American High School level.
3 A sample of 575 subjects from the target population: students, 17 years old, in their final year at Dutch VWO-schools preparing for their examinations which allow them to start academic studies. This sample was taken from two subsequent years.

Test results were analysed according to classical test theory and to item response theory. For the latter the one parameter Rasch model (Rasch, 1960) was chosen.
The Rasch model is a latent trait model. A latent trait model specifies a relationship between observable test performance and unobservable traits of abilities assumed to underlie performance on the test. The Rasch model yields estimations of the ability required to obtain a certain score on the test and of the difficulty of the items in the test on a single variable: the latent trait. The difference between person ability and item difficulty determines the probabilities of the responses of persons to items. If all items in a test measure the same ability the differences in ability amongst persons result in the same differences in probability for these persons of getting an item right for all items in the test. Similarly, the differences in mean ability of groups of persons result in equal differences in probability for these groups to succeed in each and every item in the test.
Rasch analyses were done by computer with the programme CALFIT (Wright and Mead, 1975). The unconditional maximum likelikood procedure (UCON) of the programme was used for the test data gathered from the samples from the target population. The PROX-procedure (Wright and Stone, 1979) was used to calculate item and person parameters from the data on all groups reduced randomly to 30 subjects per group to rule out influences of sample size and to compute probabilities of responses in each group.

Results

Table 1 presents mean scores and standard deviation of scores in proportion of test length and reliability (KR20) of the total listening comprehension test as observed for the three

different groups.

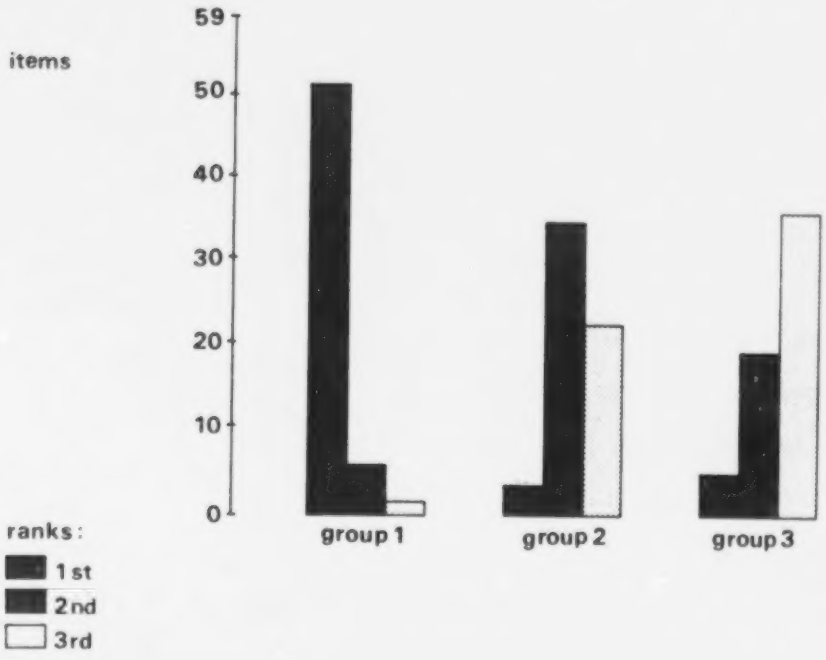
Table 1 Results of native speakers (1 and 2) and L2 Learners (3) on total test (n = 59)

Group	1	2	3
N	30	44	575
mean p	.85	.76	.72
S.D.	.05	.08	.09
KR20	.45	.56	.60

For a test of foreign language listening comprehension these results are rather disappointing, for a number of reasons. Reliability in the target group (3) is low: even at the standard length for Dutch National listening comprehension tests (75 items with two options or 50 items with three options) Spearman-Brown prediction of reliability is unacceptable (.66). Results of the second native speaker group (2) hardly differ from those in the target group (3). In fact the hypothesis that they are taken from the same population cannot be rejected (Mann-Whitney test: $p = .5$). Group 1 differs significantly from both other groups ($p < .001$), but for a group of native speakers (of comparable age and educational background as the target population) a near perfect mean score with negligible variation is to be expected if the test measures language only and at the appropriate level. According to the assumption of unidimensionality in the Rasch model, calibrations of item difficulties are population independent and therefore invariant across different groups (Hambleton and Murray, 1983). This assumption was confirmed by a correlation of .97 between UCON item calibrations calculated from two different subgroups from group 3 distinguished according to year of graduation ($N_1 = 300$, $N_2 = 275$). Correlation between PROX calibration (N = 30) and UCON calibration (N = 575) was .92. Low correlations ($\pm$ .40) of item calibrations based on the responses in the different groups show that across groups the assumption is not confirmed. Obviously item difficulty ranking differs from group to group. However, correlation of item calibrations in two subgroups of the target group (year 1 and year 2) is high (.97) which suggests some kind of bias in the test. This bias would seem due to age (group 2 is younger than group 1 and 3) and/or native language (group 1 and 2: L1, group 3: L2).
Also, if all items measure the same trait, all items should rank individuals (or groups of individuals) in the same way. Figure 1 shows that most items (88 percent) in the test consistently rank group 1 as the most able group. About two thirds rank group 2 higher than the second language learners

(3), but about one third results in a higher ranking for the second language learners.

Figure 1 Ranking of groups of native speakers and L2 learners per item



Rasch analysis (PROX-procedure) showed misfit ($p < .05$) to be unevenly distributed in the groups: misfit occurred mostly in the second native speaker group and in the group of second language learners (tabel 2) as could be expected from the data presented above.

Table 2 Mean and standard deviation of fit statistics (Z^2) total test (n = 59; N = 30)

	Group 1	Group 2	Group 3	Mean (1 + 2 + 3)
Mean	.86	1.06	1.27	1.06
S.D.	.97	1.35	1.70	.79

All items were checked for significant bias - revealed by

misfit - favouring any single group or combination of two groups: six categories in all. The data were set up in a 2 x 6 contingency table against the items selected or rejected in the previous study (De Jong, 1983). A significant relation was found between the items favouring native speakers in this study and the items constituting the best selection in the previous study (χ^2 = 31, df. 5 , p < .005), thus confirming the conclusion that two subsets of items can be distinguished each measuring a different trait. Results of the three groups in the present study on these subsets are presented in table 3.

Table 3 Results of native speakers (1 and 2) and L2 learners (3) on two subsets of items

	1	2	3
40 'best' items			
Mean p	.95	.87	.79
S.D.	.03	.07	.10
KR 20	-.15	.48	.67
19 rejected items			
Mean p	.63	.54	.58
S.D.	.11	.12	.11
KR 20	-.10	.05	-.04

The selection of 40 'best' items clearly distinguishes between the three groups (Mann Whitney test: p < .0001) and establishes the order, from high to low, group 1 - group 2 - group 3. Group 1 obtains a near perfect score and no significant variance in ability can be measured at this level amongst the individuals in this group. For the second native speaker group the selection is too easy to establish reliable differences between individuals within the group, but significant variation in scores can be observed (p < .01). In group 3, the second language learners, the test measures differences in ability best. Spearman Brown prediction for reliability at standard length is acceptable (.81), mean score is just above the ideal: midway between chance score (.5) and perfect score.

The 19 rejected items subset distinguishes less well between groups 1 and 3 (Mann Whitney test: (.01 < p < .05) and also between groups 2 and 3 (p < .01) but significant difference is observed between groups 1 and 2 (Mann Whitney test: p < .005). The order of groups 2 and 3 is reversed and group 1 remains the group with the highest scores, which suggests that difference in language ability leading to the ranking of the groups in the

40 item selection is overruled in these 19 items by a second trait in which the second language learners have higher ability. The suggestion that this trait is general intelligence and knowledge of the world (De Jong, 1983) is thus supported, group 2 being younger and having less educational background than group 3.

Rasch analyses on the responses of the three groups to the two separate subsets of items confirm the hypothesis that the trait underlying performance is different for the two subtests (table 4). In the 40-item subset all items, apart from one deviant item, test the same ability and estimate similar differences in level of ability between the three groups. The 19 rejected items fit the model less well: this subset contains three deviating items and fit statistics are relatively high considering the test is less than half the length of the 40 item selection.

Table 4 Mean and standard deviation of fit statistics (Z^2) on two subsets of items

	Group 1	Group 2	Group 3	Mean 1+2+3
40 'best' items				
Mean	.69	.63	.98	.76
S.D.	1.05	.98	1.12	1.04
19 rejected items				
Mean	.74	.67	.80	.74
S.D	.93	1.34	1.06	1.14

Rank ordering of difficulties of the items in the 40-item subset, calibrated on separate groups, is similar between all groups: correlation ranges from .92 to .98, thus confirming the hypothesis of unidimenionality of the items in this subset. Rank ordering is lower in the 19 item subset (from .70 to .79) but remains significant, suggesting that pluri-dimensionality is similarly proportioned in all items. Figure 2 presents the relevant part of the Test Characteristic Curve (TCC) for the whole test (59 items). A TCC pictures the relation between ability and probability of right answers on a test. In figure 2 the TCC is drawn only for the estimated ability of the groups in this study. An indication of the distribution of ability within the groups is given by picturing one standard deviation from the mean in both directions. The congruence of UCON-estimates, based on the responses of 575 subjects from the target population, and PROX-estimates, based

on the responses of the three groups distinguished in this study, is apparent. From the estimated means and standard deviations of ability in the three groups it is clear that, though no significant difference can be measured with this test between the second group of native speakers (2) and the group from the target population (3), group 1 stands well apart from the two other groups. Estimated variation of ability within all three groups is low: less than one logit from the 16th to the 84th percentile.

Figure 2 Test characteristic curve and distribution of ability for the total test (n = 53)

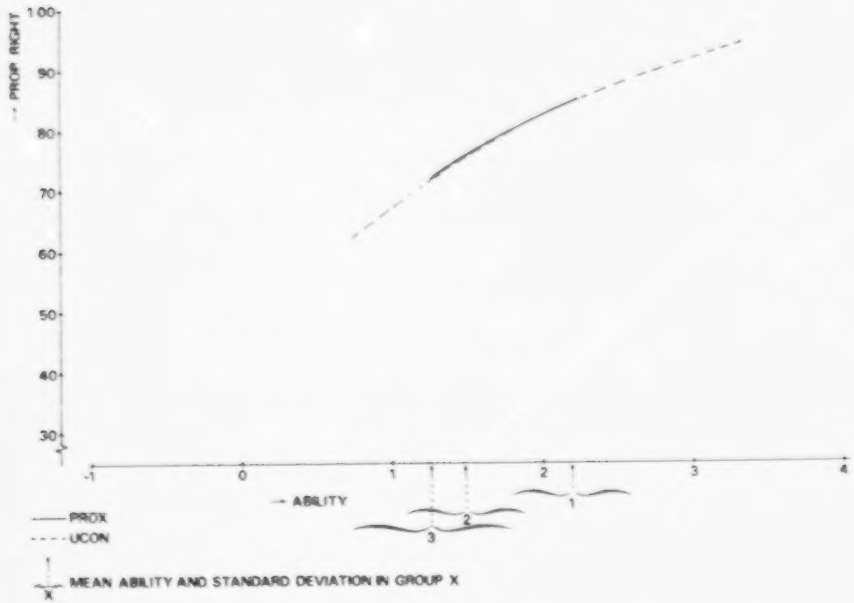


Figure 3 and 4 present TCC's for the two different subsets of items from the total test. Both figures are on the same scale as figure 2. Figure 3 shows that the 40 'best' item selection leads to an estimate of larger differences in mean ability between all three groups than the total test. However, a large amount of overlap exists between the target population (3) and the second group of native speakers (2). Because of a ceiling effect the test has no power to measure significant variation in ability amongst individuals of group 1. In the other groups there is more than one logit difference in ability between the 16th and 84th percentile.

The 19 rejected items (fig. 4) measure less than one half logit difference in ability between the means of the group lowest in

Figure 3: Test characteristic curve and distribution of ability for subset of 'best' items (n = 40)

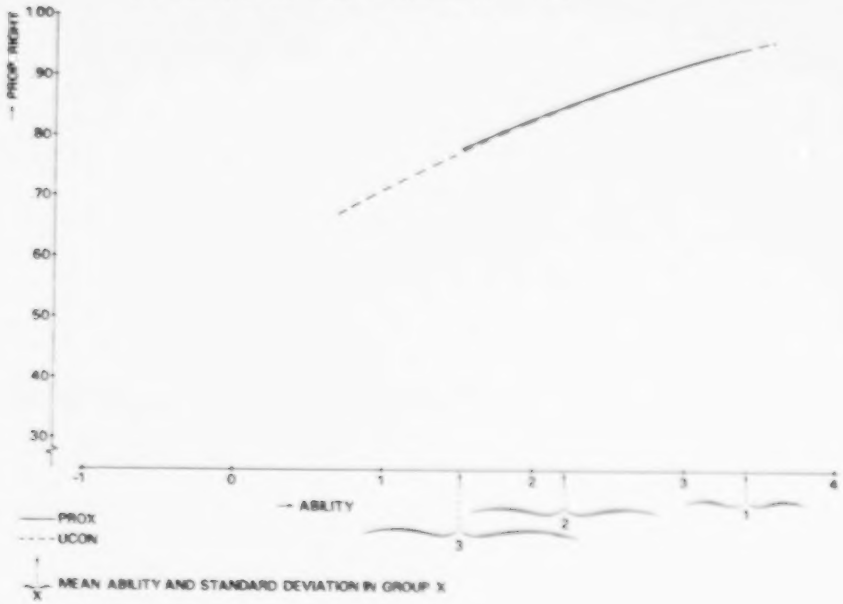
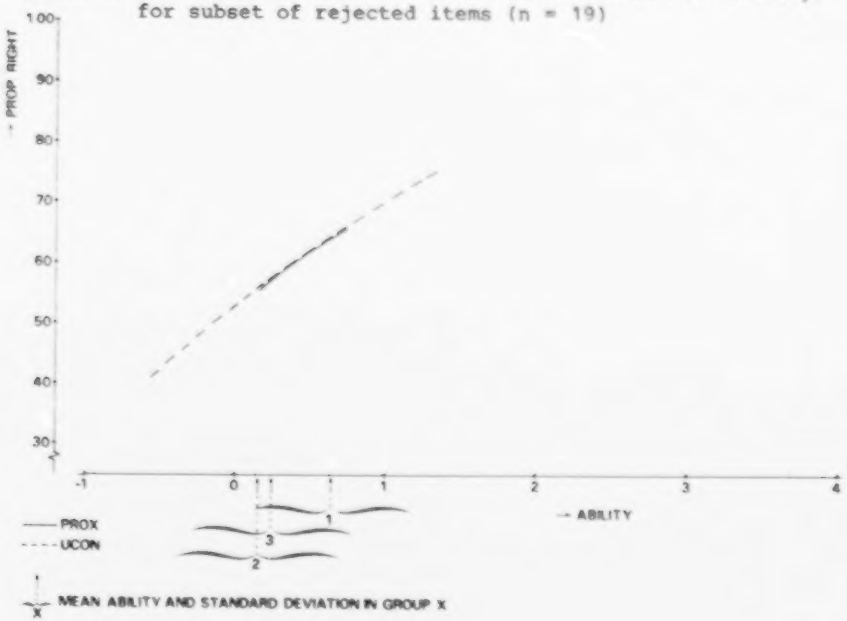


Figure 4: Test characteristic curve and distribution of ability for subset of rejected items (n = 19)



mean ability (which, on this subset, is group 21) and the group of the highest mean ability (1). Obviously there is no ceiling effect and in spite of observed scores at near chance level there is no indication of a floor effect either: guessing does not seem to have taken place. Substantial overlap between all three groups and a difference of about one logit between the 16th and 84th percentile suggest that the trait underlying this subset does discriminate but not according to any assumed difference in understanding English. Whatever trait the test measures, it is altogether different from the trait underlying the 40 'best' item selection as is indicated by the reversed position of group 2 and 3 as well as by the absence of correlation between scores of the target group on both subsets (r_{pm} = .03; N = 575).

Discussion

In a previous study (De Jong, 1983) it was concluded that a subset of 40 items from the 59 item listening comprehension test constitutes a valid measure for listening comprehension of English as a foreing language whereas 19 items had to be rejected because they measure a different ability, possibly to be identified as general intelligence or knowledge of the world. The present study demonstrates that the same selection of 40 items discriminates between two groups of native speakers differing in age and educational background and estimates significant variance in ability within the group with lower mean score. Of course L1 learners do not all achieve equally well on tests of their native language - there would be no need for L1 classes otherwise. The results of this study however suggest that language listening ability of L1 and L2 learners can be measured along a single variable and that this ability can be distinguished from an age- and school-tied variable, which could be general intelligence and/or knowledge of the world.

The groups, used in this research are small. However, Wright (1977) states that satisfactory calibrations can be achieved with tests of more than 20 items on samples of about 100 persons. Moreover, Wright and Stone (1979) successfully used a test of only 14 items on a sample of 34 subjects to demonstrate test analysis with the Rasch Model. (cf. also Lord, 1983). Wright and Stone (1979) have shown the conformity of analyses done by hand with the PROX-procedure and computer analyses with UCON. Because of significant correlation between calibrations of items on two subgroups of the target population and between calibrations on the three groups in this study, guessing cannot have seriously influenced results and calibrations apparently suffer little from error due to the small size of the groups. The short distance along the variable between Dutch students at the pre-university level and native speakers of English may be surprising at first sight. However, the level of Dutch foreign

language learners of English appears to be rather high as is clear from results on TOEFL (Test Of English As a Foreign Language) too. Clark (1977) found a mean raw score of 134.7 for native Americans, High School college-bound seniors corresponding to a scaled score of 610 (maximum 680), whereas for native Dutch L1 speakers the mean scaled score was reported to be 584 from July 1980 to June 1982, well above the mean scaled score for all participants of 503 (TOEFL, 1983).
The results reported here agree with earlier findings (Fishman, 1980; Carrel, 1980; Wilson, 1980). In the Fishman (1980) study difficulty level was artificially enhanced by adding white noise to a dictation task whereas in this study conditions were the same for all groups. Carrel (1980) studied the processing of indirectly conveyed meaning by a group of young children, native speakers of English and adults acquiring English as a second language: a much larger difference in development then the one between groups 2 and 3 in this study. Wilson (1980) could not detect first language interference even with tests purposely biased against L2 learners with elements predicted by contrastive analysis as difficult for L2 learners with a certain L1 background.
However, there is a large amount of literature revealing first language interference and language transfer (cf. Gass, 1984). Most of these studies test the hypothesis of L1 interference with discrete point tests tapping productive skills (e.g.: Schachter, 1974; Zobl, 1982, 1983; Bourgonje e.a., 1984; Van Buren and Sharwood-Smith, 1984; Van Hest e.a. 1984). Possibly, universals and language transfer operate at the receptive and productive level respectively and a combination of both principals is necessary to account for language acquisition. The claim that language teaching should begin with the receptive skills (e.g. Postovsky, 1974; Benson and Hjelt, 1980) would be consistent with Gass' suggestion (1984) that language universals serve as the overall guiding principle in language acquisition.
The present study uses an integrative test of auditory - receptive - language processing to reveal listening comprehension as a single trait for L1 and L2 learners. Language tests inevitably measure language ability on manifest behaviour: at the performance level. At the competence level it may be possible to describe language production as the reversed process of language reception. At the performance level, however, production appears to be more sensitive to language transfer than language reception is. Whether this phenomenon constitutes an intrinsic distinction between production and reception at the performance level or is only due to the fact that receptive ability - both in L1 and in L2 - is generally more developed than productive ability, remains open to further investigation.

References

Aghbar, A.A., Grid-based impressionistic scoring of ESL compositions. Paper presented at 18th. Annual TESOL convention, Toronto, 1983.

Benson, P.C. and C. Hjelt. Listening competence: a prerequisite to communication. In: Research in language testing. Eds. J.W. Oller and K. Perkins. Newbury House, Rowley, Mass., 59-65, 1980.

Bourgonje, G.C.J., P.J.M. Groot and M.M. Sharwood Smith. Universals in adverbial placement in EFL? Paper presented at the 7th World Congress of Applied Linguistics, AILA, Brussels, 1984.

Buren, P. van, and M. Sharwood Smith. Preposition-stranding as a problem for Dutch-English acquirers. Paper presented at LARS 84, University of Utrecht, 1984.

Burt, M., Dulay and S. Krashen. Language TWO. Oxford University Press, Oxford, 1982.

Canale, M., N. Frenette and M Bélanger. On the interdependence of L1 and L2 in writing: an update. Paper presented at the Sixth Pre-Tesol Language Testing Research Colloquium, Ottawa 1983.

Canale, M. and M. Swain. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics Vol 1: 1-47, 1980.

Carrel, P.L. Processing of indirectly conveyed meaning: assertion versus presupposition in first and second language acquisition. In: Research in language testing. Eds. J.W. Oller and K. Perkins. Newbury House, Rowley, Mass. 195-207, 1980.

Carroll, B.J. Testing communicative competence. Pergamon Press, Oxford, 1980.

Chomsky, N. Review: B.F. Skinner, "Verbal Behaviour". In: Language, Vol. 35, 26-58 .959.

Chomsky, N. Language and ind. Harcourt, Brace and World, New York, 1968.

Chomsky, N. Lectures on Government and binding. Foris, Dordrecht, 1981.

Clark, J.L.D. The performance of native speakers of English on the test of English as a foreign language. TOEFL Research Reports no 1, Educational Testing Service, Princeton, 1977.

Conlan, G. Comparison of analytic and holistic scoring techniques. Unpublished draft, Educational Testing Service, Princeton, N.J, 1983.

Corder, S.P. Error analysis and interlanguage. Oxford University Press, Oxford, 1981.

Corder, S.P. Review: S.D. Krashen, Second language acquisition and second language learning, Oxford 1981 and S.D. Krashen, Principles and practice in second language acquisition, Oxford 1982. In: Applied Linguistics, Vol 5: 56-58, 1984.

Fishman, M. We all make the same mistakes: a comparative study of native and nonnative errors in taking dictation. In: Research in Language Testing. Eds. J.W. Oller and K. Perkins. Newbury House, Rowley, Mass., 187-194, 1980.

Gass, S. A review of interlanguage syntax: language transfer and language universals. Language Learning Vol. 34: 115-132, 1984.

Gregg, K.R. Krashen's monitor and Occam's Razor. Applied Linguistics, Vol. 5: 79-100, 1984.

Hambleton, R.K. and L. Murray. Some goodness of fit investigations for item response models. In: Applications of item response theory. Ed. R.K. Hambleton, Educational Research Institute of British Columbia, Vancouver B.C., 71-94, 1983.

Hest, A.M. van, M.L. Kean and E. Kellerman. Some transitives transform easily. Paper presented at the 7th World Congress of Applied Linguistics, AILA, Brussels, 1984.

Jong, J.H.A.L. de. Focusing in on a latent trait: an attempt at construct validation by means of the Rasch model. In: Practice and Problems in Language Testing 5. Ed. J. van Weeren, Cito, Arnhem, 11-35, 1983.

Jong, J.H.A.L. de. Testing foreign language listening comprehension. Language Testing, Vol. 1: 97-100, 1984.

Jong, J.H.A.L. de, and H.W.M. van den Nieuwenhof. Een experimentele luistervaardigheidstoets. Specialistisch Bulletin 14, Cito, Arnhem, 1982.

Kellerman, E. Now you see it, now you don't. In: Language transfer in language learning, Eds. S. Gass and L. Selinker, Newbury House, Rowley, Mass., 1983.

Krashen, S.D., Second language acquisition and second language learning. Pergamon Press, Oxford, 1981.

Krashen, S.D., Principles and practice in second language acquisition. Pergamon Press, Oxford, 1982.

Lado, R. Linguistics across cultures. University of Michigan, Ann Arbor, 1957.

Lord, F.M. Small N justifies Rasch model. In: New horizons in testing, Ed. D.J. Weiss, Academic Press, New York, 51-61, 1983.

McLaughlin, B. The monitor model: some methodological considerations. Language Learning Vol. 28: 309-332, 1978.

Morrow, K. Communicative language testing: revolution or evolution? In: The communicative approach to language teaching. Eds. C.J. Brumfit and K. Johnson, Oxford University Press, Oxford, 1978.

Postovsky, V.A. Effects of delay in oral practice at the beginning of second language learning. Modern Language Journal Vol. 58: 229-239, 1974.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Danmarks Pedagogiske Institut, Copenhagen, 1960.

Schachter, J. An error in error analysis. Language learning 24: 205, 214, 1974.

Schachter, J. A new account of language transfer. In: Language transfer in language learning, Eds. S. Gass and L. Selinker, Newbury House, Rowley, Mass., 1983.

Spolsky, B. The limits of authenticity in language testing. Paper presented at the 7th World Congress of Applied Linguistics, AILA, Brussels, 1984.

TOEFL, test and score manual, Test of English as a Foreign Language, Princeton, N.J., 1983.

Wilson, C.B. Can ESL cloze tests be contrastively biased? - Vietnamese as a test case. In: Research in language testing. Eds. J.W. Oller and K. Perkins, Newbury House, Rowley, Mass., 208-214, 1980.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, Vol. 14: 97-117, 1977.

Wright, B.D. and R.J. Mead. CALFIT. Research memorandum 18, Department of Education, University of Chicago, 1975.

Wright, B.D. and M.H. Stone. Best test design: Rasch measurement, Mesa Press, Chicago, 1979.

Zobl, H. A direction for contrastive analysis: the comperative study of developmental sequences. TESOL. Quarterly 16: 169-184, 1982.

Zobl, H. L1 acquisition, age of L2 acquisition and the learning of word order. In: Language transfer in language learning, Eds. S. Gass and L. Selinker, Newbury House, Rowley, Mass., 1983.